明 細 書

音声合成装置

技術分野

[0001] 本発明は、合成音声を生成して出力する音声合成装置に関する。

背景技術

[0002] 従来より、所望の合成音声を生成して出力する音声合成装置が提供されている(例えば、特許文献1、特許文献2、及び特許文献3参照。)。

[0003] 特許文献1の音声合成装置は、それぞれ声質の異なる複数の音声素片データベースを備え、これらの音声素片データベースを切り替えて用いることにより、所望の合成音声を生成して出力する。

[0004] また、特許文献2の音声合成装置(音声変形装置)は、音声分析結果のスペクトルを変換することにより、所望の合成音声を生成して出力する。

[0005] また、特許文献3の音声合成装置は、複数の波形データをモーフィング処理することにより、所望の合成音声を生成して出力する。

特許文献1:特開平7－319495号公報

特許文献2:特開2000－330582号公報

特許文献3:特開平9－50295号公報

発明の開示

発明が解決しようとする課題

[0006] しかしながら、上記特許文献1及び特許文献2並びに特許文献3の音声合成装置では、声質変換の自由度が狭かったり、音質の調整が非常に困難であるという問題がある。

[0007] 即ち、特許文献1では、合成音声の声質が予め設定された声質に限られ、その予め設定された声質間の連続的な変化を表現することができない。

[0008] また、特許文献2では、スペクトルのダイナミックレンジを大きくしてしまうと音質に破綻が生じてしまい、良い音質を維持するのが困難となる。

[0009] さらに、特許文献3では、複数の波形データの互いに対応する部位(例えば波形の

ピーク)を特定して、その部位を基準にモーフィング処理を行うが、その部位を誤って特定してしまうことがある。その結果、生成された合成音声の音質が悪くなってしまう

そこで、本発明は、このような問題に鑑みてなされたものであって、声質の自由度が広く良い音質の合成音声をテキストデータから生成する音声合成装置を提供することを目的とする。

課題を解決するための手段

[0010] 上記目的を達成するために、本発明に係る音声合成装置は、第1の声質に属する複数の音声素片に関する第1の音声素片情報、及び前記第1の声質と異なる第2の声質に属する複数の音声素片に関する第2の音声素片情報を予め記憶している記憶手段と、テキストデータを取得するとともに、前記記憶手段の第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記記憶手段の第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成手段と、前記音声情報生成手段により生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段とを備え、前記音声情報生成手段は、前記第1及び第2の合成音声情報をそれぞれ複数の特徴パラメタの列として生成し、前記モーフィング手段は、前記第1及び第2の合成音声情報の互いに対応する特徴パラメタの中間値を計算することで、前記中間合成音声情報を生成することを特徴とする。

[0011] これにより、第1の声質に対する第1の音声素片情報、及び第2の声質に対する第2の音声素片情報だけを記憶手段に予め記憶させておけば、第1及び第2の声質の中間的な声質の合成音声が出力されるため、記憶手段に予め記憶させておく内容の声質に限定されずに声質の自由度を広めることができる。また、第1及び第2の声質を有する第1及び第2の合成音声情報を基礎に中間合成音声情報が生成されるため、従来例のようにスペクトルのダイナミックレンジを大きくしすぎるような処理がなさ

れず、合成音声の音質を良い状態に維持することができる。また、本発明に係る音声合成装置は、テキストデータを取得して、そこに含まれる文字列に応じた合成音声を出力するため、ユーザに対する使い勝手を向上することができる。さらに、本発明に係る音声合成装置は、第1及び第2の合成音声情報の互いに対応する特徴パラメタの中間値を計算して中間合成音声情報を生成するため、従来例のように2つのスペクトルをモーフィング処理する場合と比べて、基準とする部位を誤って特定してしまうことなく、合成音声の音質を良くすることができ、さらに、計算量を軽減することができる。

[0012] ここで、前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させることを特徴としても良い。

[0013] これにより、合成音声の出力中にその合成音声の声質が連続的に変化するため、例えば、平常声から怒り声に連続的に変化するような合成音声を出力することができる。

[0014] また、前記記憶手段は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第1及び第2の音声素片情報のそれぞれに含めて記憶しており、前記音声情報生成手段は、前記第1及び第2の合成音声情報を、それぞれに前記特徴情報を含めて生成し、前記モーフィング手段は、前記第1及び第2の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成することを特徴としても良い。例えば、前記基準は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である。また、前記音響的特徴の変化点は、前記第1及び第2の音声素片情報のそれぞれに示される各音声素片をHMM(Hidden Markov Model)で表した最尤経路上の状態遷移点であって、前記モーフィング手段は、前記第1及び第2の合成音声情報を、前記状態遷移点を用いて時間軸上で整合した上で前記中間合成音声情報を生成する。

[0015] これにより、モーフィング手段による中間合成音声情報の生成に、第1及び第2の合成音声情報が上述の基準を用いて整合されるため、例えば第1及び第2の合成音声

情報をパターンマッチングなどによって整合するような場合と比べ、迅速に整合を図って中間合成音声情報を生成することができ、その結果、処理速度を向上することができる。また、その基準をHMM(Hidden Markov Model)で表した最尤経路上の状態遷移点とすることで、第1及び第2の合成音声情報を時間軸上で正確に整合させることができる。

[0016] また、前記音声合成装置は、さらに、前記第1の声質に対応する画像を示す第1の画像情報、及び前記第2の声質に対応する画像を示す第2の画像情報を予め記憶している画像記憶手段と、前記第1及び第2の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第1及び第2の画像情報から生成する画像モーフィング手段と、前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段とを備えることを特徴としても良い。例えば、前記第1の画像情報は前記第1の声質に対応する顔画像を示し、前記第2の画像情報は前記第2の声質に対応する顔画像を示す。

[0017] これにより、第1及び第2の声質の中間的な声質に対応する顔画像が、その中間的な声質の合成音声の出力と同期して表示されるため、合成音声の声質を顔画像の表情からもユーザに伝えることができ、表現力の向上を図ることができる。

[0018] ここで、前記音声情報生成手段は、前記第1及び第2の合成音声情報のそれぞれを順次生成することを特徴としても良い。

[0019] これにより、音声情報生成手段の単位時間あたりの処理負担を軽減することができ、音声情報生成手段の構成を簡単にすることができる。その結果、装置全体を小型化することができるとともに、コスト低減を図ることができる。

[0020] また、前記音声情報生成手段は、前記第1及び第2の合成音声情報のそれぞれを並列に生成することを特徴としても良い。

[0021] これにより、第1及び第2の合成音声情報を迅速に生成することができ、その結果、テキストデータの取得から合成音声の出力までの時間を短縮することができる。

[0022] なお、本発明は、上述の音声合成装置の合成音声を生成して出力する方法やプロ

グラム、そのプログラムを格納する記憶媒体としても実現することができる。

発明の効果

[0023] 本発明の音声合成装置では、声質の自由度が広く良い音質の合成音声をテキストデータから生成することができるという効果を奏する。

図面の簡単な説明

[0024] [図1]図1は、本発明の実施の形態1に係る音声合成装置の構成を示す構成図である。

[図2]図2は、同上の音声合成部の動作を説明するための説明図である。

[図3]図3は、同上の声質指定部のディスプレイが表示する画面の一例を示す画面表示図である。

[図4]図4は、同上の声質指定部のディスプレイが表示する他の画面の一例を示す画面表示図である。

[図5]図5は、同上の音声モーフィング部の処理動作を説明するための説明図である。

[図6]図6は、同上の音声素片とHMM音素モデルの一例を示す例示図である。

[図7]図7は、同上の変形例に係る音声合成装置の構成を示す構成図である。

[図8]図8は、本発明の実施の形態2に係る音声合成装置の構成を示す構成図である。

[図9]図9は、同上の音声モーフィング部の処理動作を説明するための説明図である。

[図10]図10は、同上の声質A及び声質Zの合成音スペクトルと、それらに対応する短時間フーリエスペクトルとを示す図である。

[図11]図11は、同上のスペクトルモーフィング部が両短時間フーリエスペクトルを周波数軸上で伸縮する様子を説明するための説明図である。

[図12]図12は、同上のパワーが変換された2つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。

[図13]図13は、本発明の実施の形態3に係る音声合成装置の構成を示す構成図である。

[図14]図14は、同上の音声モーフィング部の処理動作を説明するための説明図である。

[図15]図15は、本発明の実施の形態4に係る音声合成装置の構成を示す構成図である。

[図16]図16は、同上の音声合成装置の動作を説明するための説明図である。

符号の説明

[0025] 10 テキスト

10a 音素情報

11 音声合成パラメタ値列

12 中間的合成音波形データ

12p 中間的顔画像データ

13 中間的音声合成パラメタ値列

30 音声素片

31 音素モデル

32 最尤パスの形状

41 合成音スペクトル

42 中間的合成音スペクトル

50 フォルマント形状

50a, 50b 周波数

51 フーリエスペクトル分析窓

61 合成音波形データ

101a～101z 音声合成DB

103 音声合成部

103a 言語処理部

103b 素片結合部

104 声質指定部

104A, 104B, 104Z 声質アイコン

104i 指定アイコン

105　音声モーフィング部

105a　パラメタ中間値計算部

105b　波形生成部

106　中間的合成音波形データ

107　スピーカ

203　音声合成部

201a～201z　音声合成DB

205　音声モーフィング部

205a　スペクトルモーフィング部

205b　波形生成部

303　音声合成部

301a～301z　音声合成DB

305　音声モーフィング部

305a　波形編集部

401a～401z　画像DB

405　画像モーフィング部

407　表示部

P1～P3　顔画像

発明を実施するための最良の形態

[0026] 以下、本発明の実施の形態について図面を用いて詳細に説明する。

(実施の形態1)

図1は、本発明の実施の形態1に係る音声合成装置の構成を示す構成図である。

[0027] 本実施の形態の音声合成装置は、声質の自由度が広く良い音質の合成音声をテキストデータから生成するものであって、複数の音声素片(音素)に関する音声素片データを蓄積する複数の音声合成DB101a～101zと、1つの音声合成DBに蓄積された音声素片データを用いることにより、テキスト10に示される文字列に対応する音声合成パラメタ値列11を生成する複数の音声合成部(音声情報生成手段)103と、ユーザによる操作に基づいて声質を指定する声質指定部104と、複数の音声合成

部103により生成された音声合成パラメタ値列11を用いて音声モーフィング処理を行い、中間的合成音波形データ12を出力する音声モーフィング部105と、中間的合成音波形データ12に基づいて合成音声を出力するスピーカ107とを備えている。

[0028] 音声合成DB101a～101zのそれぞれが蓄積する音声素片データの示す声質は異なっている。例えば、音声合成DB101aには、笑っている声質の音声素片データが蓄積され、音声合成DB101zには、怒っている声質の音声素片データが蓄積されている。また、本実施の形態における音声素片データは、音声生成モデルの特徴パラメタ値列の形式で表現されている。さらに、蓄積される各音声素片データには、これらのデータにより示される各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示すラベル情報が付されている。

[0029] 複数の音声合成部103は、それぞれ上述の音声合成DBと一対一に対応付けられている。このような音声合成部103の動作について図2を参照して説明する。

[0030] 図2は、音声合成部103の動作を説明するための説明図である。

音声合成部103は、図2に示すように、言語処理部103aと素片結合部103bとを備えている。

[0031] 言語処理部103aは、テキスト10を取得して、テキスト10に示される文字列を音素情報10aに変換する。音素情報10aは、テキスト10に示される文字列が音素列の形で表現されたもので、他にアクセント位置情報や音素継続長情報など、素片選択・結合・変形に必要な情報を含んでもよい。

[0032] 素片結合部103bは、対応付けられた音声合成DBの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、言語処理部103aにより出力される音素情報10aに対応する音声合成パラメタ値列11を生成する。音声合成パラメタ値列11は、実際の音声波形を生成するために必要となる十分な情報を含んだ複数の特徴パラメタの値が配列されたものである。例えば、音声合成パラメタ値列11は、時系列に沿った各音声分析合成フレームごとに、図2に示すような、5つの特徴パラメタを含んで構成される。5つの特徴パラメタとは、音声の基本周波数F0と、第一フォルマントF1と、第二フォルマントF2と、音声分析合成フレーム継続長FRと、音源強度PWとである。また、上述のように音声素片データには

ラベル情報が付されているので、このように生成される音声合成パラメタ値列11にもラベル情報が付されている。

[0033] 声質指定部104は、ユーザによる操作に基づき、何れの音声合成パラメタ値列11を用い、その音声合成パラメタ値列11に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部105に指示する。さらに、声質指定部104はその割合を時系列に沿って変化させる。このような声質指定部104は、例えばパーソナルコンピュータなどから構成され、ユーザにより操作された結果を表示するディスプレイを備えている。

[0034] 図3は、声質指定部104のディスプレイが表示する画面の一例を示す画面表示図である。

[0035] ディスプレイには、音声合成DB101a～101zの声質を示す複数の声質アイコンが表示されている。なお図3では、複数の声質アイコンのうち、声質Aの声質アイコン104Aと、声質Bの声質アイコン104Bと、声質Zの声質アイコン104Zとを示す。このような複数の声質アイコンは、それぞれの示す声質が似ているものほど互いに近寄るように配置され、似ていないものほど互いに離れるように配置される。

[0036] ここで、声質指定部104は、このようなディスプレイ上に、ユーザによる操作に応じて移動可能な指定アイコン104iを表示する。

[0037] 声質指定部104は、ユーザによって配置された指定アイコン104iから近い声質アイコンを調べ、例えば声質アイコン104A, 104B, 104Zを特定すると、声質Aの音声合成パラメタ値列11と、声質Bの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを用いることを、音声モーフィング部105に指示する。さらに、声質指定部104は、各声質アイコン104A, 104B, 104Z及び指定アイコン104iの相対的な配置に対応する割合を、音声モーフィング部105に指示する。

[0038] 即ち、声質指定部104は、指定アイコン104iから各声質アイコン104A, 104B, 104Zまでの距離を調べ、それらの距離に応じた割合を指示する。

[0039] 又は、声質指定部104は、まず、声質Aと声質Zの中間的な声質（テンポラリ声質）を生成するための割合を求め、次に、そのテンポラリ声質と声質Bとから、指定アイコン104iで示される声質を生成するための割合を求め、これらの割合を指示する。具

体的に、声質指定部104は、声質アイコン104A及び声質アイコン104Zを結ぶ直線と、声質アイコン104B及び指定アイコン104iを結ぶ直線とを算出し、これらの直線の交点の位置104tを特定する。この位置104tにより示される声質が上述のテンポラリ声質である。そして、声質指定部104は、位置104tから各声質アイコン104A，104Zまでの距離の割合を求める。次に、声質指定部104は、指定アイコン104iから声質アイコン104B及び位置104tまでの距離の割合を求め、このように求めた2つの割合を指示する。

[0040] このような声質指定部104を操作することにより、ユーザは、スピーカ107から出力させようとする合成音声の声質の、予め設定された声質に対する類似度を容易に入力することができる。そこでユーザは、例えば声質Aに近い合成音声をスピーカ107から出力させたいときには、指定アイコン104iが声質アイコン104Aに近づくように声質指定部104を操作する。

[0041] また、声質指定部104は、ユーザからの操作に応じて、上述のような割合を時系列に沿って連続的に変化させる。

[0042] 図4は、声質指定部104のディスプレイが表示する他の画面の一例を示す画面表示図である。

[0043] 声質指定部104は、図4に示すように、ユーザによる操作に応じて、ディスプレイ上に3つのアイコン21，22，23を配置し、アイコン21からアイコン22を通ってアイコン23に到達するような軌跡を特定する。そして、声質指定部104は、その軌跡に沿って指定アイコン104iが移動するように、上述の割合を時系列に沿って連続的に変化させる。例えば、声質指定部104は、その軌跡の長さをLとすると、毎秒0.01×Lの速度で指定アイコン104iが移動するように、その割合を変化させる。

[0044] 音声モーフィング部105は、上述のような声質指定部104により指定された音声合成パラメタ値列11と割合とから、音声モーフィング処理を行う。

[0045] 図5は、音声モーフィング部105の処理動作を説明するための説明図である。

音声モーフィング部105は、図5に示すように、パラメタ中間値計算部105aと、波形生成部105bとを備えている。

[0046] パラメタ中間値計算部105aは、声質指定部104により指定された少なくとも2つの

音声合成パラメタ値列11と割合とを特定し、それらの音声合成パラメタ値列11から、互いに対応する音声分析合成フレーム間ごとに、その割合に応じた中間的音声合成パラメタ値列13を生成する。

[0047] 例えば、パラメタ中間値計算部105aは、声質指定部104の指定に基づいて、声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11と、割合50:50とを特定すると、まず、その声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを、それぞれに対応する音声合成部103から取得する。そして、パラメタ中間値計算部105aは、互いに対応する音声分析合成フレームにおいて、声質Aの音声合成パラメタ値列11に含まれる各特徴パラメタと、声質Zの音声合成パラメタ値列11に含まれる各特徴パラメタとの中間値を50:50の割合で算出し、その算出結果を中間的音声合成パラメタ値列13として生成する。具体的に、互いに対応する音声分析合成フレームにおいて、声質Aの音声合成パラメタ値列11の基本周波数F0の値が300であり、声質Zの音声合成パラメタ値列11の基本周波数F0の値が280である場合には、パラメタ中間値計算部105aは、当該音声分析合成フレームでの基本周波数F0が290となる中間的音声合成パラメタ値列13を生成する。

[0048] また、図3を用いて説明したように、声質指定部104により、声質Aの音声合成パラメタ値列11と、声質Bの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とが指定され、さらに、声質Aと声質Zの中間的なテンポラリ声質を生成するための割合(例えば3:7)と、そのテンポラリ声質と声質Bとから指定アイコン104iで示される声質を生成するための割合(例えば9:1)とが指定され場合には、音声モーフィング部105は、まず、声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを用いて、3:7の割合に応じた音声モーフィング処理を行う。これにより、テンポラリ声質に対応する音声合成パラメタ値列が生成される。さらに、音声モーフィング部105は、先に生成した音声合成パラメタ値列と、声質Bの音声合成パラメタ値列11とを用いて、9:1の割合に応じた音声モーフィング処理を行う。これにより、指定アイコン104iに対応する中間的音声合成パラメタ値列13が生成される。ここで、上述の3:7の割合に応じた音声モーフィング処理とは、声質Aの音声合成パラメタ値列11を3/(3+7)だけ声質Zの音声合成パラメタ値列11に近づける処理であり、逆に、声質

Zの音声合成パラメタ値列11を7／(3＋7)だけ声質Aの音声合成パラメタ値列11に近づける処理をいう。この結果、生成される音声合成パラメタ値列は、声質Zの音声合成パラメタ値列11よりも、声質Aの音声合成パラメタ値列11に類似することとなる。

[0049] 波形生成部105bは、パラメタ中間値計算部105aにより生成された中間的音声合成パラメタ値列13を取得して、その中間的音声合成パラメタ値列13に応じた中間的合成音波形データ12を生成し、スピーカ107に対して出力する。

[0050] これにより、スピーカ107からは、中間的音声合成パラメタ値列13に応じた合成音声が出力される。即ち、予め設定された複数の声質の中間的な声質の合成音声がスピーカ107から出力される。

[0051] ここで、一般に複数の音声合成パラメタ値列11に含まれる音声分析合成フレームの総数はそれぞれ異なるため、パラメタ中間値計算部105aは、上述のように互いに異なる声質の音声合成パラメタ値列11を用いて音声モーフィング処理を行うときには、音声分析合成フレーム間の対応付けを行うために時間軸アライメントを行う。

[0052] 即ちパラメタ中間値計算部105aは、音声合成パラメタ値列11に付されたラベル情報に基づいて、これらの音声合成パラメタ値列11の時間軸上の整合を図る。

[0053] ラベル情報は、前述のように各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示す。音響的特徴の変化点は、例えば、音声素片に対応する不特定話者HMM音素モデルにより示される最尤パスの状態遷移点である。

[0054] 図6は、音声素片とHMM音素モデルの一例を示す例示図である。

例えば、図6に示すように、所定の音声素片30を不特定話者HMM音素モデル(以下、音素モデルと略す)31で認識した場合、その音素モデル31は、開始状態(S_0)と終了状態(S_E)を含めて4つの状態(S_0, S_1, S_2, S_E)で構成される。ここで、最尤パスの形状32は、時刻4から5において、状態S1から状態S2への状態遷移を有する。つまり、音声合成DB101a～101zに格納されている音声素片データの音声素片30に対応する部分には、この音声素片30の開始時刻1、終了時刻N、及び音響的特徴の変化点の時刻5を示すラベル情報が付されている。

[0055] したがって、パラメタ中間値計算部105aは、そのラベル情報に示される開始時刻1、終了時刻N、及び音響的特徴の変換点の時刻5に基づいて、時間軸の伸縮処理を

行う。即ち、パラメタ中間値計算部105aは、取得した各音声合成パラメタ値列11に対して、ラベル情報により示される時刻が一致するように、その時刻間を線形に伸縮する。

[0056] これにより、パラメタ中間値計算部105aは、各音声合成パラメタ値列11に対して、それぞれの音声分析合成フレームの対応付けを行うことができる。つまり、時間軸アライメントを行うことができる。また、このように本実施の形態ではラベル情報を用いて時間軸アライメントを行うことにより、例えば各音声合成パラメタ値列11のパターンマッチングなどにより時間軸アライメントを行う場合と比べて、迅速に時間軸アライメントを実行することができる。

[0057] 以上のように本実施の形態では、パラメタ中間値計算部105aが、声質指定部104から指示された複数の音声合成パラメタ値列11に対して、声質指定部104から指定された割合に応じた音声モーフィング処理を実行するため、合成音声の声質の自由度を広めることができる。

[0058] 例えば、図3に示す声質指定部104のディスプレイ上で、ユーザが声質指定部104を操作することにより指定アイコン104iを声質アイコン104A、声質アイコン104B及び声質アイコン104Zに近づければ、音声モーフィング部105は、声質Aの音声合成DB101aに基づいて音声合成部103により生成された音声合成パラメタ値列11と、声質Bの音声合成DB101bに基づいて音声合成部103により生成された音声合成パラメタ値列11と、声質Zの音声合成DB101zに基づいて音声合成部103により生成された音声合成パラメタ値列11とを用いて、それぞれを同じ割合で音声モーフィング処理する。その結果、スピーカ107から出力される合成音声を、声質Aと声質Bと声質Cとの中間的な声質にすることができる。また、ユーザが声質指定部104を操作することにより指定アイコン104iを声質アイコン104Aに近づければ、スピーカ107から出力される合成音声の声質を声質Aに近づけることができる。

[0059] また、本実施の形態の声質指定部104は、ユーザによる操作に応じてその割合を時系列に沿って変化させるため、スピーカ107から出力される合成音声の声質を時系列に沿ってなめらかに変化させることができる。例えば、図4で説明したように、声質指定部104が、毎秒0.01×Lの速度で軌跡上を指定アイコン104iが移動するよう

に割合を変化させた場合には、100秒間声質がなめらかに変化し続けるような合成音声がスピーカ107から出力される。

[0060] これによって、例えば「喋り始めは冷静だが、喋りながら段々怒っていく」というような、従来は不可能だった、表現力の高い音声合成装置が実現できる。また、合成音声の声質を1発声の中で連続的に変化させることもできる。

[0061] さらに、本実施の形態では、音声モーフィング処理を行うため、従来例のように声質に破綻が起こることがなく合成音声の品質を維持することができる。また、本実施の形態では、声質の異なる音声合成パラメタ値列11の互いに対応する特徴パラメタの中間値を計算して中間的音声合成パラメタ値列13を生成するため、従来例のように2つのスペクトルをモーフィング処理する場合と比べて、基準とする部位を誤って特定してしまうことなく、合成音声の音質を良くすることができ、さらに、計算量を軽減することができる。また、本実施の形態では、HMMの状態遷移点を用いることで、複数の音声合成パラメタ値列11を時間軸上で正確に整合させることができる。即ち、声質Aの音素の中でも、状態遷移点を基準に前半と後半とで音響的特徴が異なり、声質Bの音素の中でも、状態遷移点を基準に前半と後半とで音響的特徴が異なる場合がある。このような場合に、声質Aの音素と声質Bの音素とをそれぞれ単純に時間軸に伸縮して、それぞれの発声時間を合わせても、つまり時間軸アライメントを行っても、両音素からモーフィング処理された音素には、各音素の前半と後半とが入り乱れてしまう。しかし、上述のようにHMMの状態遷移点を用いると、各音素の前半と後半とが入り乱れてしまうのを防ぐことができる。その結果、モーフィング処理された音素の音質を良くして、所望の中間的な声質の合成音声を出力することができる。

[0062] なお、本実施の形態では、複数の音声合成部103のそれぞれに音素情報10a及び音声合成パラメタ値列11を生成させたが、音声モーフィング処理に必要となる声質に対応する音素情報10aが何れも同じであるときには、1つの音声合成部103の言語処理部103aにのみ音素情報10aを生成させ、その音素情報10aから音声合成パラメタ値列11を生成する処理を、複数の音声合成部103の素片結合部103bにさせても良い。

[0063] （変形例）

ここで、本実施の形態における音声合成部に関する変形例について説明する。

[0064] 図7は、本変形例に係る音声合成装置の構成を示す構成図である。

本変形例に係る音声合成装置は、互いに異なる声質の音声合成パラメタ値列11を生成する1つの音声合成部103cを備える。

[0065] この音声合成部103cは、テキスト10を取得して、テキスト10に示される文字列を音素情報10aに変換した後、複数の音声合成DB101a～101zを順番に切り替えて参照ことで、その音素情報10aに対応する複数の声質の音声合成パラメタ値列11を順次生成する。

[0066] 音声モーフィング部105は、必要な音声合成パラメタ値列11が生成されるまで待機し、その後、上述と同様の方法で中間的合成音波形データ12を生成する。

[0067] なお、上述のような場合、声質指定部104は、音声合成部103cに指示して、音声モーフィング部105が必要とする音声合成パラメタ値列11のみを生成させることで、音声モーフィング部105の待機時間を短くすることができる。

[0068] このように本変形例では、音声合成部103cを1つだけ備えることにより、音声合成装置全体の小型化並びにコスト低減を図ることができる。

[0069] （実施の形態2）

図8は、本発明の実施の形態2に係る音声合成装置の構成を示す構成図である。

[0070] 本実施の形態の音声合成装置は、実施の形態1の音声合成パラメタ値列11の代わりに周波数スペクトルを用い、この周波数スペクトルによる音声モーフィング処理を行う。

[0071] このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成DB201a～201zと、1つの音声合成DBに蓄積された音声素片データを用いることにより、テキスト10に示される文字列に対応する合成音スペクトル41を生成する複数の音声合成部203と、ユーザによる操作に基づいて声質を指定する声質指定部104と、複数の音声合成部203により生成された合成音スペクトル41を用いて音声モーフィング処理を行い、中間的合成音波形データ12を出力する音声モーフィング部205と、中間的合成音波形データ12に基づいて合成音声を出力するスピーカ107とを備えている。

[0072] 複数の音声合成DB201a～201zのそれぞれが蓄積する音声素片データの示す声質は、実施の形態1の音声合成DB101a～101zと同様、異っている。また、本実施の形態における音声素片データは、周波数スペクトルの形式で表現されている。

[0073] 複数の音声合成部203は、それぞれ上述の音声合成DBと一対一に対応付けられている。そして、各音声合成部203は、テキスト10を取得して、テキスト10に示される文字列を音素情報に変換する。さらに、音声合成部203は、対応付けられた音声合成DBの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する周波数スペクトルたる合成音スペクトル41を生成する。このような合成音スペクトル41は、音声のフーリエ解析結果の形式であっても良く、音声のケプストラムパラメタ値を時系列的に並べた形式であっても良い。

[0074] 声質指定部104は、実施の形態1と同様、ユーザによる操作に基づき、何れの合成音スペクトル41を用い、その合成音スペクトル41に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部205に指示する。さらに、声質指定部104はその割合を時系列に沿って変化させる。

[0075] 本実施の形態における音声モーフィング部205は、複数の音声合成部203から出力される合成音スペクトル41を取得して、その中間的性質を持つ合成音スペクトルを生成し、さらに、その中間的性質の合成音スペクトルを中間的合成音波形データ12に変形して出力する。

[0076] 図9は、本実施の形態における音声モーフィング部205の処理動作を説明するための説明図である。

[0077] 音声モーフィング部205は、図9に示すように、スペクトルモーフィング部205aと、波形生成部205bとを備えている。

[0078] スペクトルモーフィング部205aは、声質指定部104により指定された少なくとも2つの合成音スペクトル41と割合とを特定し、それらの合成音スペクトル41から、その割合に応じた中間的合成音スペクトル42を生成する。

[0079] 即ち、スペクトルモーフィング部205aは、複数の合成音スペクトル41から、声質指定部104により指定された2つ以上の合成音スペクトル41を選択する。そして、スペ

クトルモーフィング部205aは、それら合成音スペクトル41の形状の特徴を示すフォルマント形状50を抽出して、そのフォルマント形状50ができるだけ一致するような変形を各合成音スペクトル41に加えた後、各合成音スペクトル41の重ね合わせを行う。なお、上述の合成音スペクトル41の形状の特徴は、フォルマント形状でなくても良く、例えばある程度以上強く現れていて、かつその軌跡が連続的に追えるものであれば良い。図9に示されるように、フォルマント形状50は、声質Aの合成音スペクトル41及び声質Zの合成音スペクトル41のそれぞれについてスペクトル形状の特徴を模式的に表すものである。

[0080] 具体的に、スペクトルモーフィング部205aは、声質指定部104からの指定に基づき、声質A及び声質Zの合成音スペクトル41と4:6の割合とを特定すると、まず、その声質Aの合成音スペクトル41と声質Zの合成音スペクトル41とを取得して、それらの合成音スペクトル41からフォルマント形状50を抽出する。次に、スペクトルモーフィング部205aは、声質Aの合成音スペクトル41のフォルマント形状50が声質Zの合成音スペクトル41のフォルマント形状50に40%だけ近づくように、声質Aの合成音スペクトル41を周波数軸及び時間軸上で伸縮処理する。さらに、スペクトルモーフィング部205aは、声質Zの合成音スペクトル41のフォルマント形状50が声質Aの合成音スペクトル41のフォルマント形状50に60%だけ近づくように、声質Zの合成音スペクトル41を周波数軸及び時間軸上で伸縮処理する。最後に、スペクトルモーフィング部205aは、伸縮処理された声質Aの合成音スペクトル41のパワーを60%にするとともに、伸縮処理された声質Zの合成音スペクトル41のパワーを40%にした上で、両合成音スペクトル41を重ね合わせる。その結果、声質Aの合成音スペクトル41と声質Zの合成音スペクトル41との音声モーフィング処理が4:6の割合で行われ、中間的合成音スペクトル42が生成される。

[0081] このような、中間的合成音スペクトル42を生成する音声モーフィング処理について、図10～図12を用いてより詳細に説明する。

[0082] 図10は、声質A及び声質Zの合成音スペクトル41と、それらに対応する短時間フーリエスペクトルとを示す図である。

[0083] スペクトルモーフィング部205aは、声質Aの合成音スペクトル41と声質Zの合成音

スペクトル41との音声モーフィング処理を4:6の割合で行うときには、まず、上述のようにこれらの合成音スペクトル41のフォルマント形状50を互いに近づけるため、各合成音スペクトル41同士の時間軸アライメントを行う。このような時間軸アライメントは、各合成音スペクトル41のフォルマント形状50同士のパターンマッチングを行うことにより実現される。なお、各合成音スペクトル41もしくはフォルマント形状50に関する他の特徴量を用いてパターンマッチングを行ってもよい。

[0084] 即ち、スペクトルモーフィング部205aは、図10に示すように、両合成音スペクトル41のそれぞれのフォルマント形状50において、パターンが一致するフーリエスペクトル分析窓51の部位で時刻が一致するように、両合成音スペクトル41に対して時間軸上の伸縮を行う。これにより時間軸アライメントが実現される。

[0085] また、図10に示すように、互いにパターンが一致するフーリエスペクトル分析窓51のそれぞれの短時間フーリエスペクトル41aには、フォルマント形状50の周波数50a，50bが互いに異なるように表示される。

[0086] そこで、時間軸アライメントの完了後、スペクトルモーフィング部205aは、アライメントされた音声の各時刻において、フォルマント形状50を基に、周波数軸上の伸縮処理を行う。即ち、スペクトルモーフィング部205aは、各時刻における声質A及び声質Bの短時間フーリエスペクトル41aにおいて周波数50a，50bが一致するように、両短時間フーリエスペクトル41aを周波数軸上で伸縮する。

[0087] 図11は、スペクトルモーフィング部205aが両短時間フーリエスペクトル41aを周波数軸上で伸縮する様子を説明するための説明図である。

[0088] スペクトルモーフィング部205aは、声質Aの短時間フーリエスペクトル41a上の周波数50a，50bが40%だけ、声質Zの短時間フーリエスペクトル41a上の周波数50a，50bに近付くように、声質Aの短時間フーリエスペクトル41aを周波数軸上で伸縮し、中間的な短時間フーリエスペクトル41bを生成する。これと同様に、スペクトルモーフィング部205aは、声質Zの短時間フーリエスペクトル41a上の周波数50a，50bが60%だけ、声質Aの短時間フーリエスペクトル41a上の周波数50a，50bに近付くように、声質Zの短時間フーリエスペクトル41aを周波数軸上で伸縮し、中間的な短時間フーリエスペクトル41bを生成する。その結果、中間的な両短時間フーリエスペクトル

41bにおいて、フォルマント形状50の周波数は周波数f1，f2に揃えられた状態となる。

[0089] 例えば、声質Aの短時間フーリエスペクトル41a上でフォルマント形状50の周波数50a，50bが500Hz及び3000Hzであり、声質Zの短時間フーリエスペクトル41a上でフォルマント形状50の周波数50a，50bが400Hz及び4000Hzであり、かつ各合成音のナイキスト周波数が11025Hzである場合を想定して説明する。スペクトルモーフィング部205aは、まず、声質Aの短時間フーリエスペクトル41aの帯域f＝0～500Hzが0～(500＋(400－500)×0．4)Hzとなるように、帯域f＝500～3000Hzが(500＋(400－500)×0．4)～(3000＋(4000－3000)×0．4)Hzとなるように、帯域f＝3000～11025Hzが(3000＋(4000－3000)×0．4)～11025Hzとなるように、声質Aの短時間フーリエスペクトル41aに対して周波数軸上の伸縮・移動を行う。これと同様に、スペクトルモーフィング部205aは、声質Zの短時間フーリエスペクトル41aの帯域f＝0～400Hzが0～(400＋(500－400)×0．6)Hzとなるように、帯域f＝400～4000Hzが(400＋(500－400)×0．6)～(4000＋(3000－4000)×0．6)Hzとなるように、帯域f＝4000～11025Hzが(4000＋(3000－4000)×0．6)～11025Hzとなるように、声質Zの短時間フーリエスペクトル41aに対して周波数軸上の伸縮・移動を行う。その伸縮・移動の結果により生成された2つの短時間フーリエスペクトル41bにおいて、フォルマント形状50の周波数は周波数f1，f2に揃えられた状態となる。

[0090] 次に、スペクトルモーフィング部205aは、このような周波数軸上の変形が行われた両短時間フーリエスペクトル41bのパワーを変形する。即ち、スペクトルモーフィング部205aは、声質Aの短時間フーリエスペクトル41bのパワーを60%に変換し、声質Zの短時間フーリエスペクトル41bのパワーを40%に変換する。そして、スペクトルモーフィング部205aは、上述のように、パワーが変換されたこれらの短時間フーリエスペクトルを重ね合わせる。

[0091] 図12は、パワーが変換された2つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。

[0092] この図12に示すように、スペクトルモーフィング部205aは、パワーが変換された声

質Aの短時間フーリエスペクトル41cと、同じくパワーが変換された声質Bの短時間フーリエスペクトル41cとを重ね合わせ、新たな短時間フーリエスペクトル41dを生成する。このとき、スペクトルモーフィング部205aは、互いの短時間フーリエスペクトル41cの上記周波数f1，f2を一致させた状態で、両短時間フーリエスペクトル41cを重ね合わせる。

[0093] そして、スペクトルモーフィング部205aは、上述のような短時間フーリエスペクトル41dの生成を、両合成音スペクトル41の時間軸アライメントされた時刻ごとに行う。その結果、声質Aの合成音スペクトル41と声質Zの合成音スペクトル41との音声モーフィング処理が4:6の割合で行われ、中間的合成音スペクトル42が生成されるのである。

[0094] 音声モーフィング部205の波形生成部205bは、上述のようにスペクトルモーフィング部205aにより生成された中間的合成音スペクトル42を、中間的合成音波形データ12に変換して、これをスピーカ107に出力する。その結果、スピーカ107から、中間的合成音スペクトル42に対応する合成音声が出力される。

[0095] このように、本実施の形態においても、実施の形態1と同様、声質の自由度が広く良い音質の合成音声をテキスト10から生成することができる。

[0096] （変形例）

ここで、本実施の形態におけるスペクトルモーフィング部の動作に関する変形例について説明する。

[0097] 本変形例に係るスペクトルモーフィング部は、上述のように合成音スペクトル41からその形状の特徴を示すフォルマント形状50を抽出して用いることなく、音声合成DBに予め格納されたスプライン曲線の制御点の位置を読み出して、そのスプライン曲線をフォルマント形状50の代わりに用いる。

[0098] 即ち、各音声素片に対応するフォルマント形状50を、周波数対時間の2次元平面上の複数のスプライン曲線と見なし、そのスプライン曲線の制御点の位置を予め音声合成DBに格納しておく。

[0099] このように、本変形例に係るスペクトルモーフィング部は、合成音スペクトル41からわざわざフォルマント形状50を抽出することをせず、音声合成DBに予め格納されて

いる制御点の位置が示すスプライン曲線を用いて時間軸及び周波数軸上の変換処理を行うため、上記変換処理を迅速に行うことができる。

[0100] なお、上述のようなスプライン曲線の制御点の位置ではなくフォルマント形状50そのものを、予め音声合成DB201a～201zに格納しておいても良い。

[0101] （実施の形態3）

図13は、本発明の実施の形態3に係る音声合成装置の構成を示す構成図である。

[0102] 本実施の形態の音声合成装置は、実施の形態1の音声合成パラメタ値列11や、実施の形態2の合成音スペクトル41の代わりに音声波形を用い、この音声波形による音声モーフィング処理を行う。

[0103] このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成DB301a～301zと、1つの音声合成DBに蓄積された音声素片データを用いることにより、テキスト10に示される文字列に対応する合成音波形データ61を生成する複数の音声合成部303と、ユーザによる操作に基づいて声質を指定する声質指定部104と、複数の音声合成部303により生成された合成音波形データ61を用いて音声モーフィング処理を行い、中間的合成音波形データ12を出力する音声モーフィング部305と、中間的合成音波形データ12に基づいて合成音声を出力するスピーカ107とを備えている。

[0104] 複数の音声合成DB301a～301zのそれぞれが蓄積する音声素片データの示す声質は、実施の形態1の音声合成DB101a～101zと同様、異なっている。また、本実施の形態における音声素片データは、音声波形の形式で表現されている。

[0105] 複数の音声合成部303は、それぞれ上述の音声合成DBと一対一に対応付けられている。そして、各音声合成部303は、テキスト10を取得して、テキスト10に示される文字列を音素情報に変換する。さらに、音声合成部303は、対応付けられた音声合成DBの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する音声波形たる合成音波形データ61を生成する。

[0106] 声質指定部104は、実施の形態1と同様、ユーザによる操作に基づき、何れの合成音波形データ61を用い、その合成音波形データ61に対してどのような割合で音声

モーフィング処理を行うかを音声モーフィング部305に指示する。さらに、声質指定部104はその割合を時系列に沿って変化させる。

[0107] 本実施の形態における音声モーフィング部305は、複数の音声合成部303から出力される合成音波形データ61を取得して、その中間的性質を持つ中間的合成音波形データ12を生成して出力する。

[0108] 図14は、本実施の形態における音声モーフィング部305の処理動作を説明するための説明図である。

[0109] 本実施の形態における音声モーフィング部305は波形編集部305aを備えている。この波形編集部305aは、声質指定部104により指定された少なくとも2つの合成音波形データ61と割合とを特定し、それらの合成音波形データ61から、その割合に応じた中間的合成音波形データ12を生成する。

[0110] 即ち、波形編集部305aは、複数の合成音波形データ61から、声質指定部104により指定された2つ以上の合成音波形データ61を選択する。そして、波形編集部305aは、声質指定部104により指定された割合に応じ、その選択した合成音波形データ61のそれぞれに対して、例えば各音声の各サンプリング時点におけるピッチ周波数や振幅、各音声における各有声区間の継続時間長などを変形する。波形編集部305aは、そのように変形された合成音波形データ61を重ね合わせることで、中間的合成音波形データ12を生成する。

[0111] スピーカ107は、このように生成された中間的合成音波形データ12を波形編集部305aから取得して、その中間的合成音波形データ12に対応する合成音声を出力する。

[0112] このように、本実施の形態においても、実施の形態1又は2と同様、声質の自由度が広く良い音質の合成音声をテキスト10から生成することができる。

[0113] (実施の形態4)

図15は、本発明の実施の形態4に係る音声合成装置の構成を示す構成図である。

[0114] 本実施の形態の音声合成装置は、出力する合成音声の声質に応じた顔画像を表示するものであって、実施の形態1に含まれる構成要素と、複数の顔画像に関する画像情報を蓄積する複数の画像DB401a～401zと、これらの画像DB401a～401zに

蓄積される顔画像の情報を用いて画像モーフィング処理を行い、中間的顔画像データ12pを出力する画像モーフィング部405と、画像モーフィング部405から中間的顔画像データ12pを取得して、その中間的顔画像データ12pに応じた顔画像を表示する表示部407とを備えている。

[0115] 画像DB401a～401zのそれぞれが蓄積する画像情報の示す顔画像の表情は異なっている。例えば、怒っている声質の音声合成DB101aに対応する画像DB401aには、怒っている表情の顔画像に関する画像情報が蓄積されている。また、画像DB401a～401zに蓄積されている顔画像の画像情報には、顔画像の眉及び口の端や中央、目の中心点など、この顔画像の表す表情の印象をコントロールするための特徴点が付加されている。

[0116] 画像モーフィング部405は、声質指定部104により指定された各合成音声パラメタ値列102のそれぞれの声質に対応付けされた画像DBから画像情報を取得する。そして、画像モーフィング部405は、取得した画像情報を用いて、声質指定部104により指定された割合に応じた画像モーフィング処理を行う。

[0117] 具体的に、画像モーフィング部405は、取得した一方の画像情報により示される顔画像の特徴点の位置が、声質指定部104により指定された割合だけ、取得した他方の画像情報により示される顔画像の特徴点の位置に変位するように、その一方の顔画像をワーピングし、これと同様に、その他方の顔画像の特徴点の位置を、声質指定部104により指定された割合だけ、その一方の顔画像の特徴点の位置に変位するように、その他方の顔画像をワーピングする。そして、画像モーフィング部405は、ワーピングされたそれぞれの顔画像を、声質指定部104により指定された割合に応じてクロスディゾルブすることで、中間的顔画像データ12pを生成する。

[0118] これにより本実施の形態では、例えばエージェントの顔画像と合成音声の声質の印象を常に一致させることができる。即ち、本実施の形態の音声合成装置は、エージェントの平常声と怒り声の間の音声モーフィングを行って、少しだけ怒った声質の合成音声を生成するときには、音声モーフィングと同様の比率でエージェントの平常顔画像と怒り顔画像の間の画像モーフィングを行い、エージェントのその合成音声に適した少しだけ怒った顔画像を表示する。言い換えれば、感情を持つエージェントに対し

てユーザが感じる聴覚的印象と、視覚的印象を一致させることができ、エージェントの提示する情報の自然性を高めることができる。

[0119] 図16は、本実施の形態の音声合成装置の動作を説明するための説明図である。

例えば、ユーザが声質指定部104を操作することにより、図3に示すディスプレイ上の指定アイコン104iを、声質アイコン104Aと声質アイコン104Zを結ぶ線分を4:6に分割する位置に配置すると、音声合成装置は、スピーカ107から出力される合成音声が10%だけ声質A寄りになるように、その4:6の割合に応じた音声モーフィング処理を声質A及び声質Zの音声合成パラメタ値列11を用いて行い、声質A及び声質Bの中間的な声質xの合成音声を出力する。これと同時に、音声合成装置は、上記割合と同じ4:6の割合に応じた画像モーフィング処理を、声質Aに対応付けられた顔画像P1と、声質Zに対応付けられた顔画像P2とを用いて行い、これらの画像の中間的な顔画像P3を生成して表示する。ここで、音声合成装置は、画像モーフィングするときには、上述のように、顔画像P1の眉や口の端などの特徴点の位置を、顔画像P2の眉や口の端などの特徴点の位置に向けて40%の割合で変化するように、その顔画像P1をワーピングし、これと同様に、顔画像P2の特徴点の位置を、顔画像P1の特徴点の位置に向けて60%の割合で変化するように、その顔画像P2をワーピングする。そして、画像モーフィング部405は、ワーピングされた顔画像P1に対して60%の割合で、ワーピングされた顔画像P2に対して40%の割合でクロスディゾルブし、その結果、顔画像P3を生成する。

[0120] このように、本実施の形態の音声合成装置は、スピーカ107から出力する合成音声の声質が「怒っている」ときには、「怒っている」様子の顔画像を表示部407に表示し、声質が「泣いている」ときには、「泣いている」様子の顔画像を表示部407に表示する。さらに、本実施形態の音声合成装置は、その声質が「怒っている」ものと「泣いている」ものとの中間的なものであるときには、「怒っている」顔画像と「泣いている」顔画像の中間的な顔画像を表示するとともに、その声質が「怒っている」ものから「泣いている」ものへと時間的に変化するときには、中間的な顔画像をその声質に一致させて時間的に変化させる。

[0121] なお、画像モーフィングは他にも様々な方法によって可能であるが、元となる画像

の間の比率を指定することで目的の画像が指定できる方法であれば、どんなものを用いてもよい。

産業上の利用可能性

[0122] 本発明は、声質の自由度が広く良い音質の合成音声をテキストデータから生成することができるという効果を有し、ユーザに対して感情を表す合成音声を出力する音声合成装置などに適用することができる。

請求の範囲

[1] 第1の声質に属する複数の音声素片に関する第1の音声素片情報、及び前記第1の声質と異なる第2の声質に属する複数の音声素片に関する第2の音声素片情報を予め記憶している記憶手段と、

テキストデータを取得するとともに、前記記憶手段の第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記記憶手段の第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成手段と、

前記音声情報生成手段により生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、

前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段と

を備え、

前記音声情報生成手段は、前記第1及び第2の合成音声情報をそれぞれ複数の特徴パラメタの列として生成し、

前記モーフィング手段は、前記第1及び第2の合成音声情報の互いに対応する特徴パラメタの中間値を計算することで、前記中間合成音声情報を生成する

ことを特徴とする音声合成装置。

[2] 前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる

ことを特徴とする請求項1記載の音声合成装置。

[3] 前記記憶手段は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第1及び第2の音声素片情報のそれぞれに含めて記憶しており、

前記音声情報生成手段は、前記第1及び第2の合成音声情報を、それぞれに前記

特徴情報を含めて生成し、

前記モーフィング手段は、前記第1及び第2の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項1記載の音声合成装置。

[4] 前記基準は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である

ことを特徴とする請求項3記載の音声合成装置。

[5] 前記音響的特徴の変化点は、前記第1及び第2の音声素片情報のそれぞれに示される各音声素片をHMM(Hidden Markov Model)で表した最尤経路上の状態遷移点であって、

前記モーフィング手段は、前記第1及び第2の合成音声情報を、前記状態遷移点を用いて時間軸上で整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項4記載の音声合成装置。

[6] 前記音声合成装置は、さらに、

前記第1の声質に対応する画像を示す第1の画像情報、及び前記第2の声質に対応する画像を示す第2の画像情報を予め記憶している画像記憶手段と、

前記第1及び第2の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第1及び第2の画像情報から生成する画像モーフィング手段と、

前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段と

を備えることを特徴とする請求項1記載の音声合成装置。

[7] 前記第1の画像情報は前記第1の声質に対応する顔画像を示し、前記第2の画像情報は前記第2の声質に対応する顔画像を示す

ことを特徴とする請求項6記載の音声合成装置。

[8] 前記音声合成装置は、さらに、

前記第1及び第2の声質を示す固定点、及びユーザの操作に基づいて移動する移動点をそれぞれN次元（Nは自然数）の座標上に配置して表し、前記固定点及び移動点の配置に基づいて、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を導出し、導出した割合を前記モーフィング手段に指示する指定手段を備え、

前記モーフィング手段は、前記指定手段により指定された割合に応じて、前記中間合成音声情報を生成する

ことを特徴とする請求項1記載の音声合成装置。

[9] 前記音声情報生成手段は、

前記第1及び第2の合成音声情報のそれぞれを順次生成する

ことを特徴とする請求項1記載の音声合成装置。

[10] 前記音声情報生成手段は、

前記第1及び第2の合成音声情報のそれぞれを並列に生成する

ことを特徴とする請求項1記載の音声合成装置。

[11] 第1の声質に属する複数の音声素片に関する第1の音声素片情報、及び前記第1の声質と異なる第2の声質に属する複数の音声素片に関する第2の音声素片情報を予め記憶しているメモリを用いることで、合成音声を生成して出力する音声合成方法であって、

テキストデータを取得するテキスト取得ステップと、

前記メモリの第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記メモリの第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成ステップと、

前記音声情報生成ステップで生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィングステップと、

前記モーフィングステップで生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力ステップと

を含み、

前記音声情報生成ステップでは、前記第1及び第2の合成音声情報をそれぞれ複数の特徴パラメタの列として生成し、

前記モーフィングステップでは、前記第1及び第2の合成音声情報の互いに対応する特徴パラメタの中間値を計算することで、前記中間合成音声情報を生成する

ことを特徴とする音声合成方法。

[12] 前記モーフィングステップでは、前記音声出力ステップで出力される合成音声の声質がその出力中に連続的に変化するように、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる

ことを特徴とする請求項11記載の音声合成方法。

[13] 前記メモリは、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第1及び第2の音声素片情報のそれぞれに含めて記憶しており、

前記音声情報生成ステップでは、前記第1及び第2の合成音声情報を、それぞれに前記特徴情報を含めて生成し、

前記モーフィングステップでは、前記第1及び第2の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項11記載の音声合成方法。

[14] 前記基準は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である

ことを特徴とする請求項13記載の音声合成方法。

[15] 前記音響的特徴の変化点は、前記第1及び第2の音声素片情報のそれぞれに示される各音声素片をHMM(Hidden Markov Model)で表した最尤経路上の状態遷移点であって、

前記モーフィングステップでは、前記第1及び第2の合成音声情報を、前記状態遷移点を用いて時間軸上で整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項14記載の音声合成方法。

[16] 前記音声合成方法は、さらに、

前記第1の声質に対応する画像を示す第1の画像情報、及び前記第2の声質に対応する画像を示す第2の画像情報を予め記憶している画像メモリを用い、

前記第1及び第2の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記画像メモリの第1及び第2の画像情報から生成する画像モーフィングステップと、

前記画像モーフィングステップで生成された中間画像情報により示される画像を、前記音声出力ステップで出力される合成音声に同期させて表示する表示ステップと

を含むことを特徴とする請求項11記載の音声合成方法。

[17] 前記第1の画像情報は前記第1の声質に対応する顔画像を示し、前記第2の画像情報は前記第2の声質に対応する顔画像を示す

ことを特徴とする請求項16記載の音声合成方法。

[18] 第1の声質に属する複数の音声素片に関する第1の音声素片情報、及び前記第1の声質と異なる第2の声質に属する複数の音声素片に関する第2の音声素片情報を予め記憶しているメモリを用いることで、合成音声を生成して出力するためのプログラムであって、

テキストデータを取得するテキスト取得ステップと、

前記メモリの第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記メモリの第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成ステップと、

前記音声情報生成ステップで生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィングステップと、

前記モーフィングステップで生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力ステップと

をコンピュータに実行させ、

前記音声情報生成ステップでは、前記第1及び第2の合成音声情報をそれぞれ複

数の特徴パラメタの列として生成し、

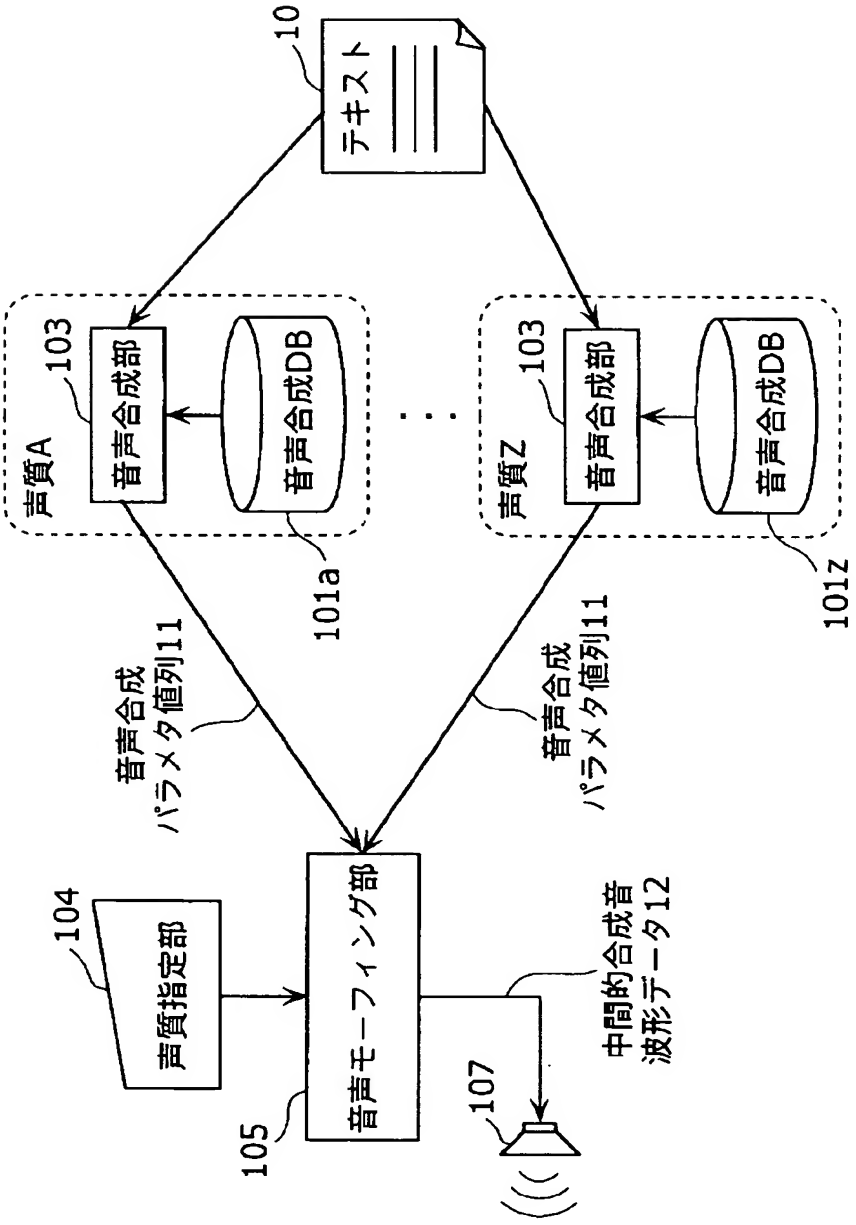
前記モーフィングステップでは、前記第1及び第2の合成音声情報の互いに対応する特徴パラメタの中間値を計算することで、前記中間合成音声情報を生成することを特徴とするプログラム。

要 約 書

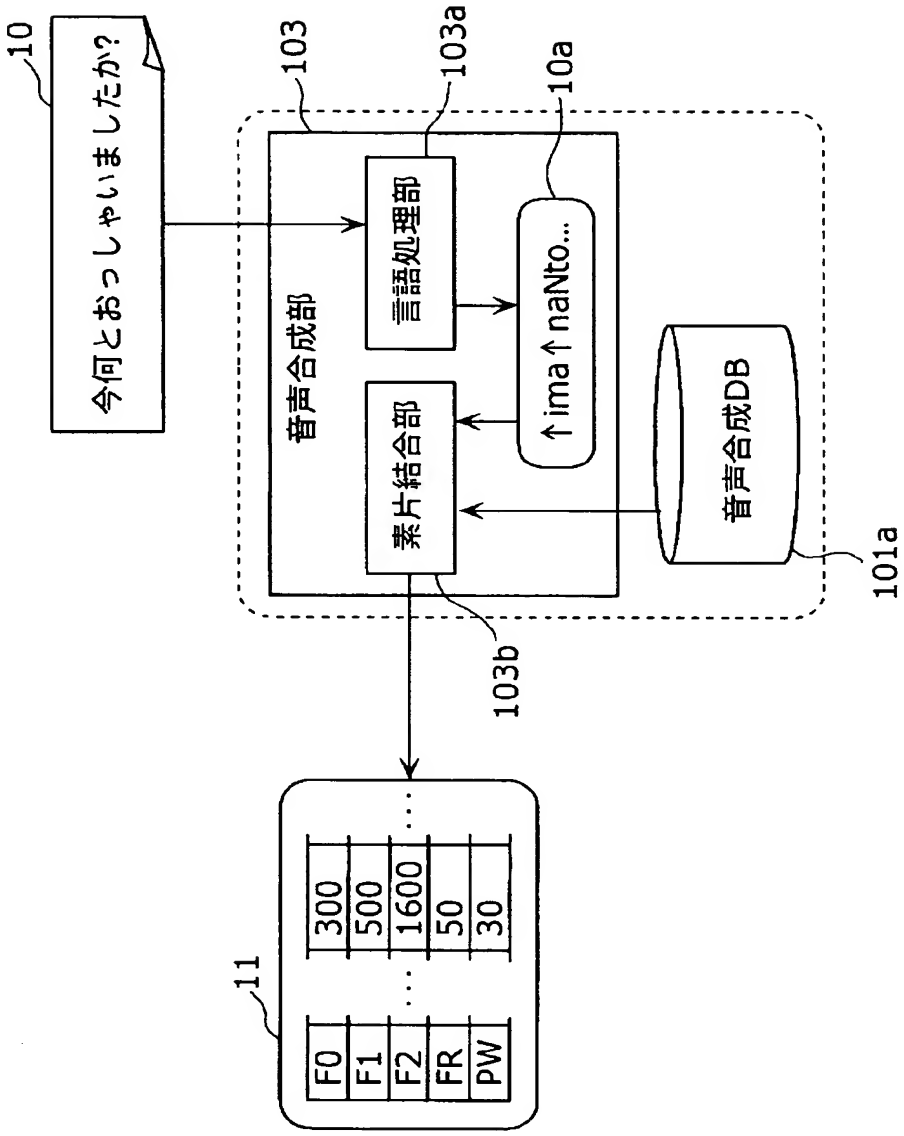
声質の自由度が広く良い音質の合成音声をテキストデータから生成する音声合成装置を提供する。

音声合成装置は、音声合成DB(101a, 101z)と、テキスト(10)を取得するとともに、音声合成DB(101a)から、テキスト(10)に含まれる文字に対応した声質Aの音声合成パラメタ値列(11)を生成する音声合成部(103)と、音声合成DB(101z)から、テキスト(10)に含まれる文字に対応した声質Zの音声合成パラメタ値列(11)を生成する音声合成部(103)と、声質A及び声質Zの音声合成パラメタ値列(11)から、テキスト(10)に含まれる文字に対応した、声質A及び声質Zの中間的な声質の合成音声を示す中間的音声合成パラメタ値列(13)を生成する音声モーフィング部(105)と、生成された中間的音声合成パラメタ値列(13)をその合成音声に変換して出力するスピーカ(107)とを備える。

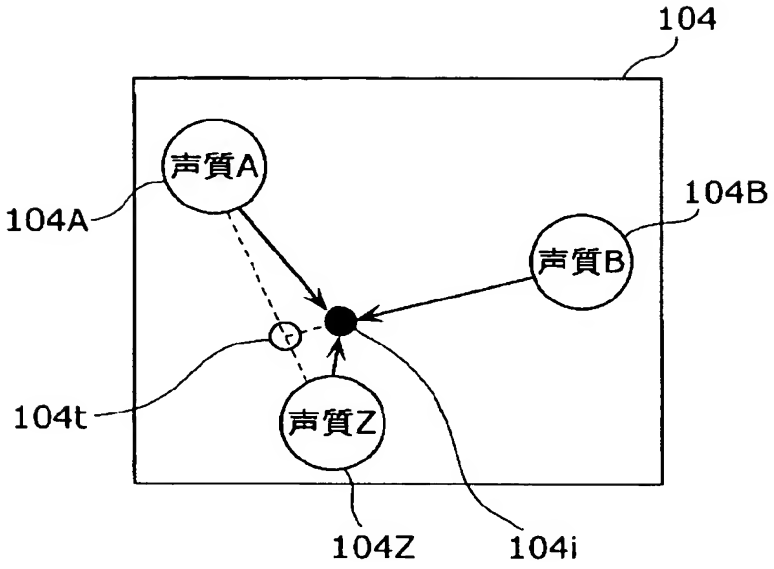
[図1]


10
テキスト
声質A
103
音声合成部
音声合成DB
101a
声質Z
103
音声合成部
音声合成DB
101z
音声合成
パラメタ値列11
音声合成
パラメタ値列11
104
声質指定部
音声モーフィング部
105
107
中間的合成音
波形データ12

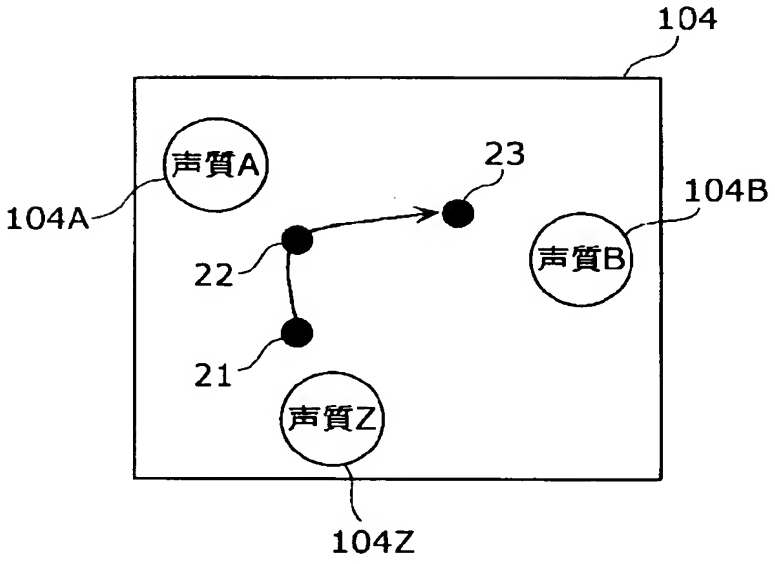
[図2]


10
今何とおっしゃいましたか?
103
103a
10a
音声合成部
言語処理部
素片結合部
↑ima↑naNto...
音声合成DB
101a
103b
11
F0
F1
F2
FR
PW
300
500
1600
50
30

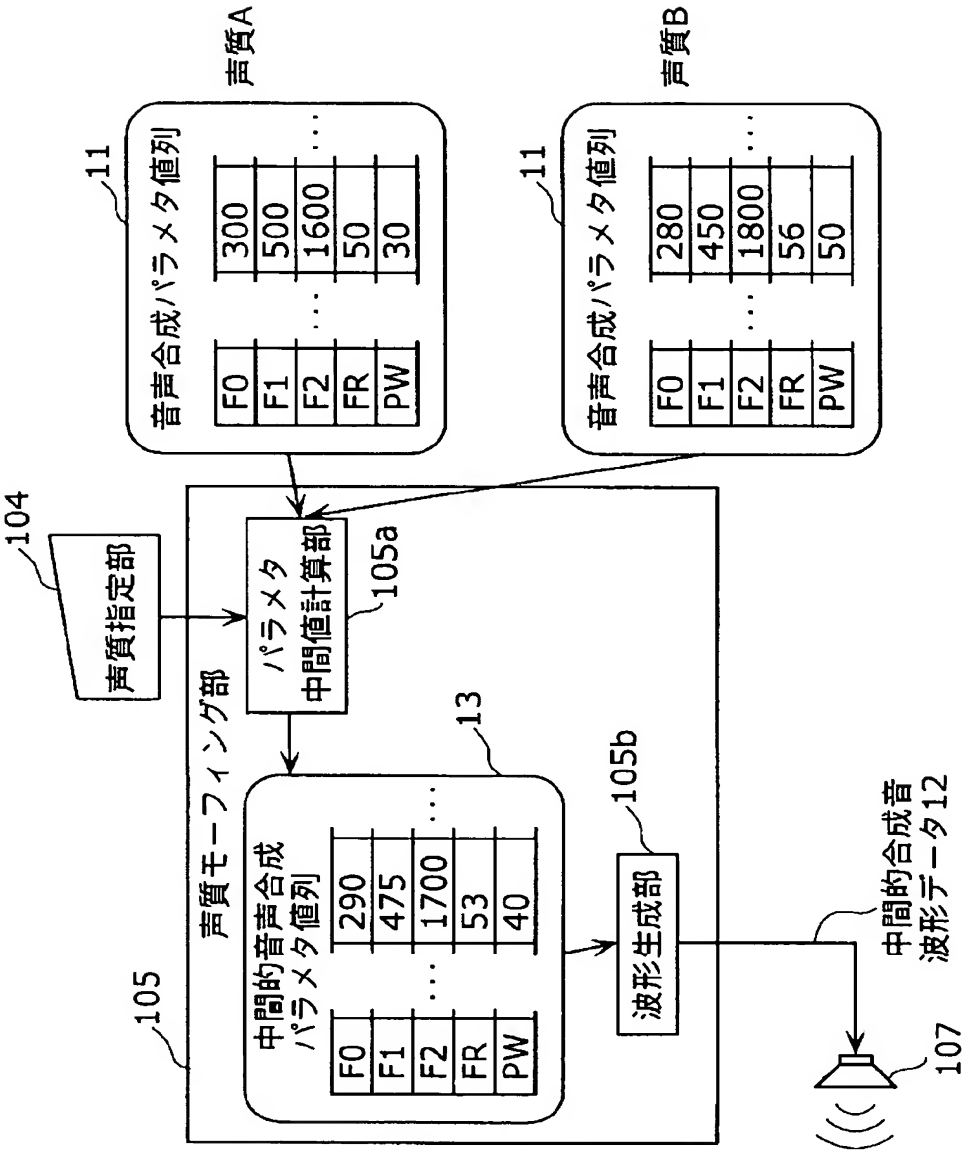
[図3]


104
声質A
104A
104B
声質B
104t
声質Z
104Z
104i

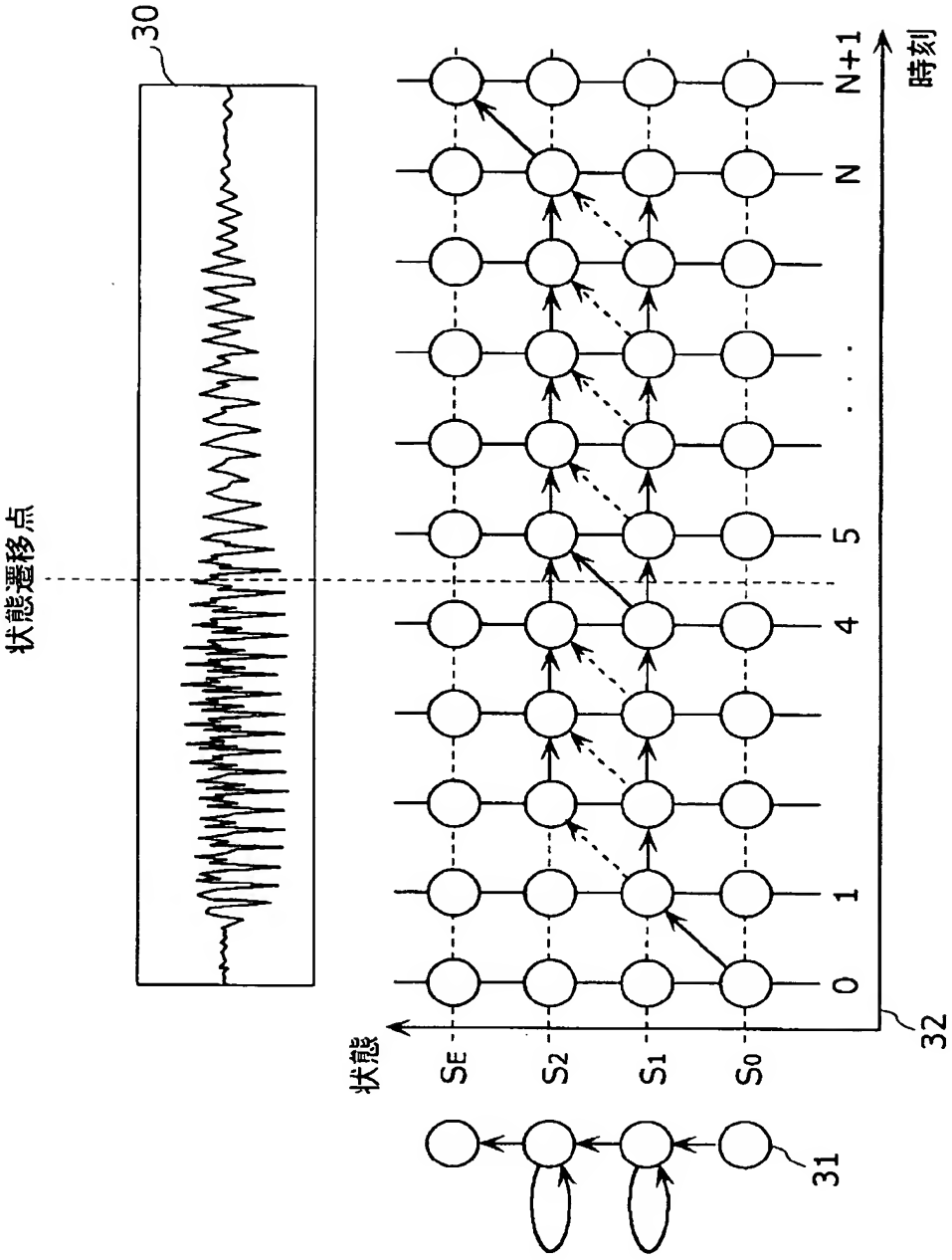
[図4]


104
声質A
23
104A
104B
声質B
22
21
声質Z
104Z

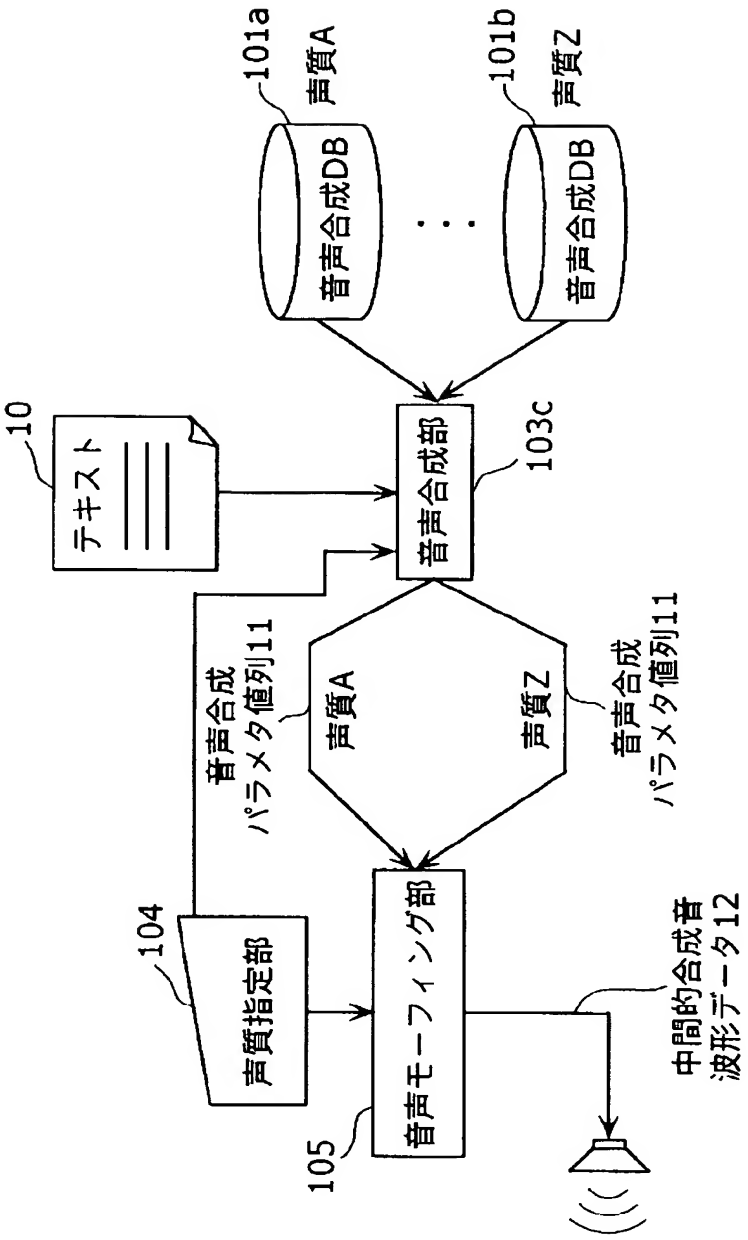
[図5]


声質A
11
音声合成パラメタ値列
F0
F1
F2
FR
PW
300
500
1600
50
30
声質B
11
音声合成パラメタ値列
F0
F1
F2
FR
PW
280
450
1800
56
50
104
声質指定部
パラメタ
中間値計算部
105a
声質モーフィング部
13
中間的音声合成
パラメタ値列
F0
F1
F2
FR
PW
290
475
1700
53
40
105b
波形生成部
105
中間的合成音
波形データ12
107

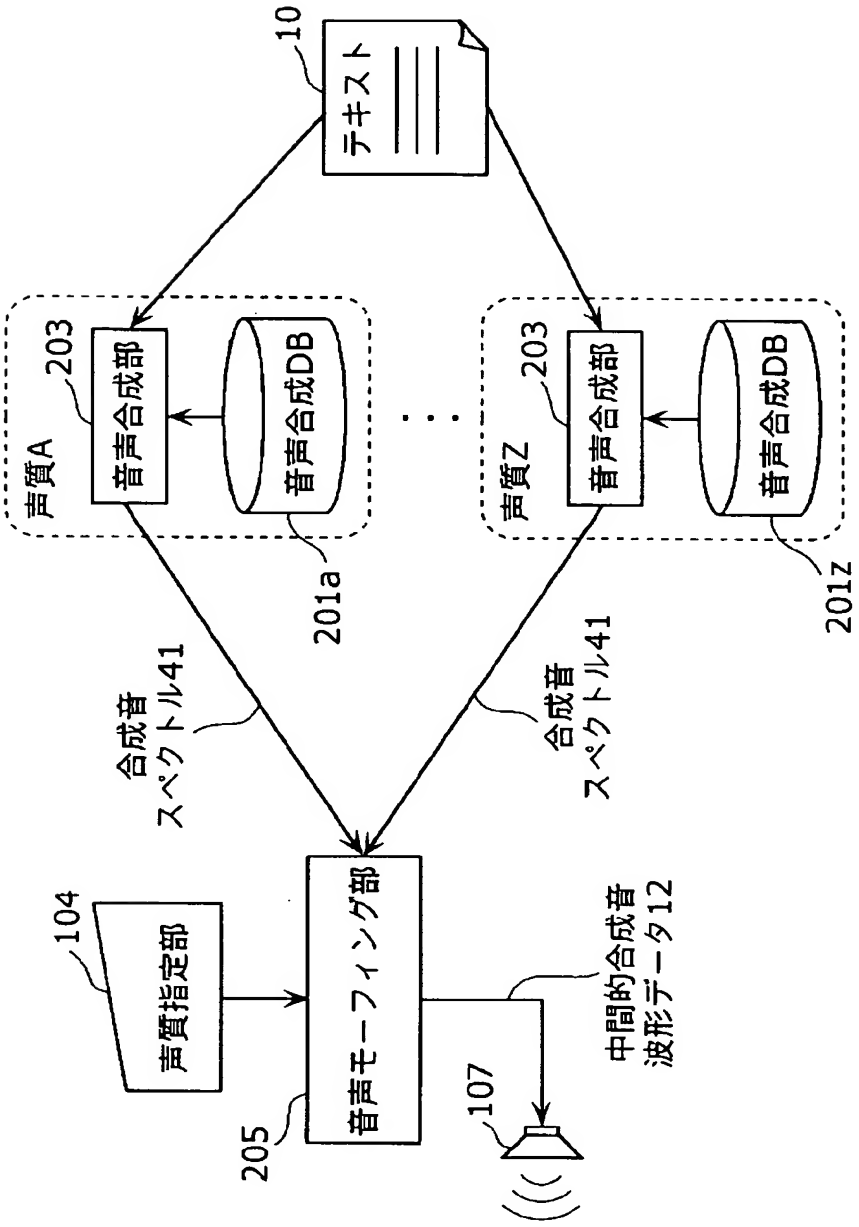
[図6]



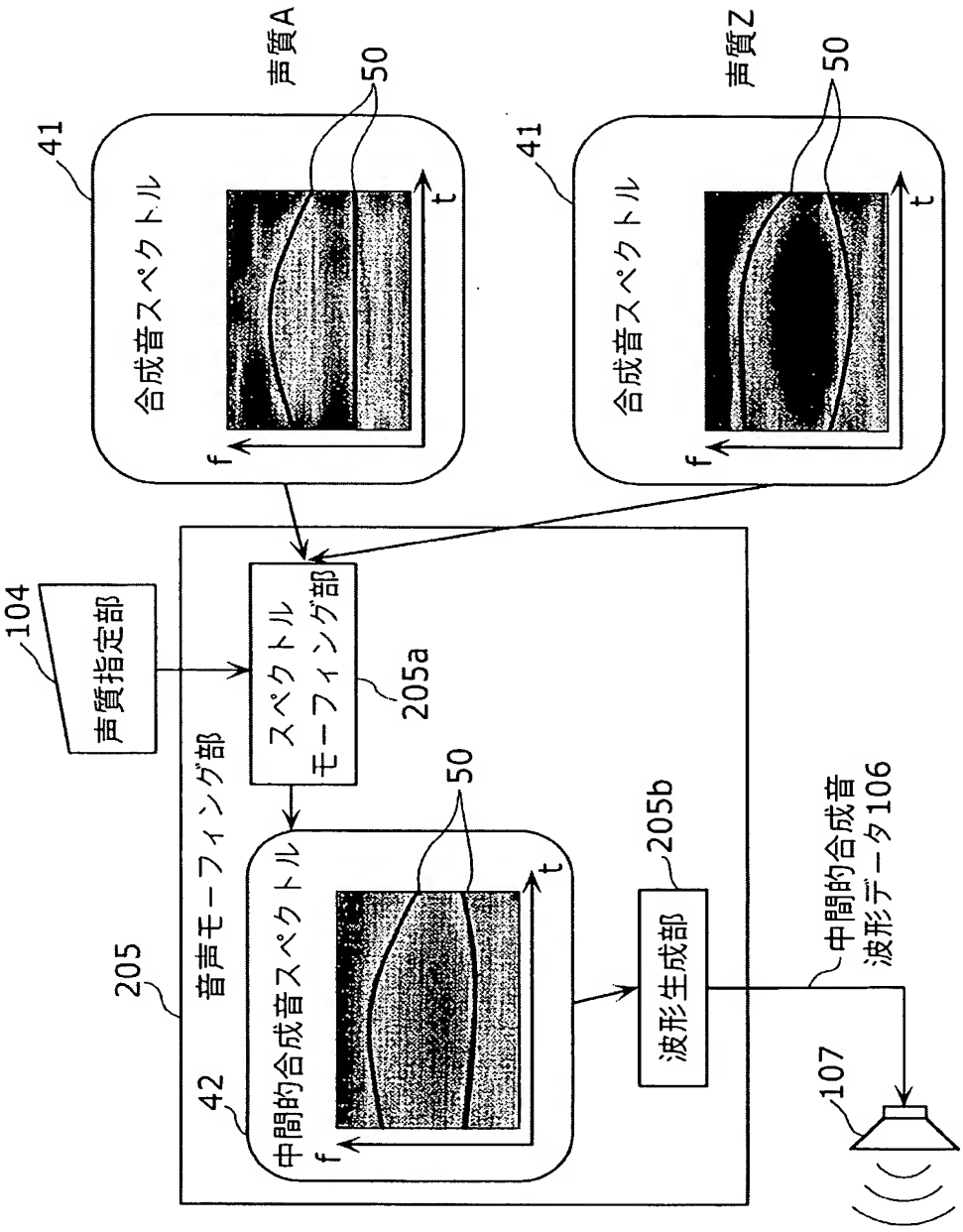
[図7]


101a
声質A
音声合成DB
101b
声質Z
音声合成DB
10
テキスト
音声合成部
103c
音声合成
パラメタ値列11
声質A
声質Z
音声合成
パラメタ値列11
104
声質指定部
音声モーフィング部
105
中間的合成音
波形データ12

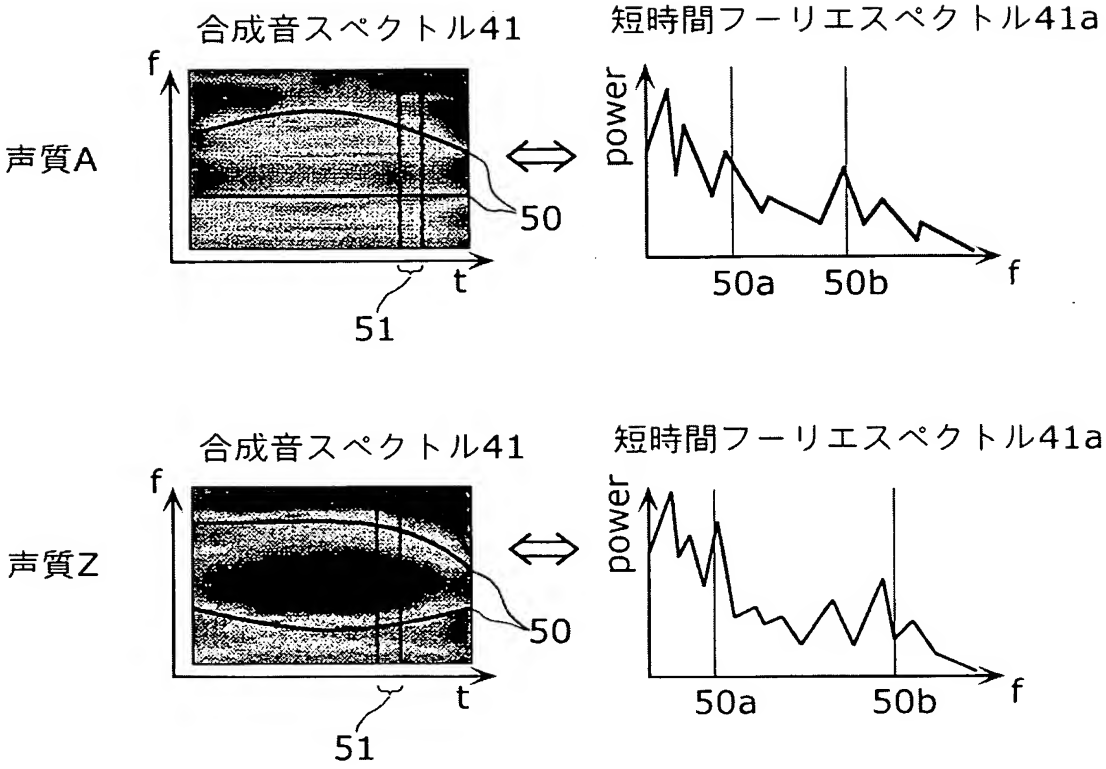
[図8]


10
テキスト
声質A
203
音声合成部
音声合成DB
201a
声質Z
203
音声合成部
音声合成DB
201z
合成音
スペクトル41
合成音
スペクトル41
104
声質指定部
音声モーフィング部
205
107
中間的合成音
波形データ12

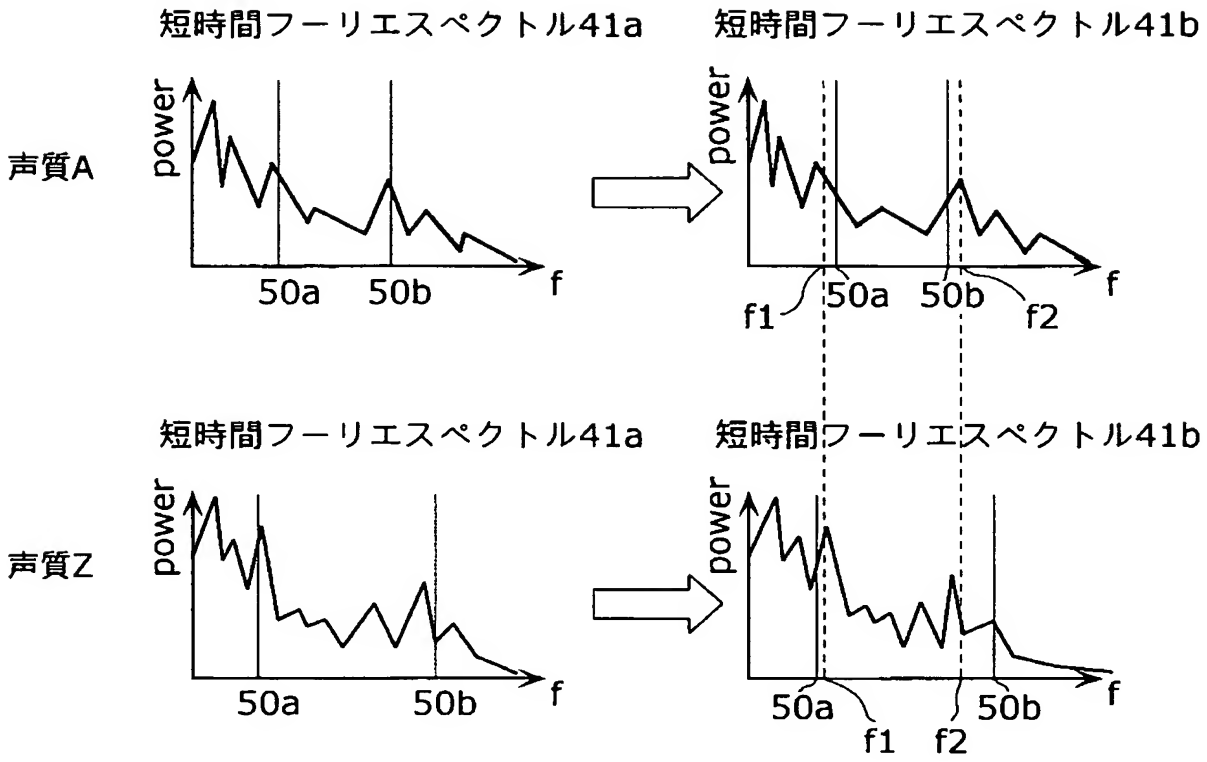
[図9]


声質A
50
41
合成音スペクトル
t
f
声質Z
50
41
合成音スペクトル
t
f
104
声質指定部
スペクトル
モーフィング部
205a
205
音声モーフィング部
50
42
中間的合成音スペクトル
t
f
205b
波形生成部
中間的合成音
波形データ106
107

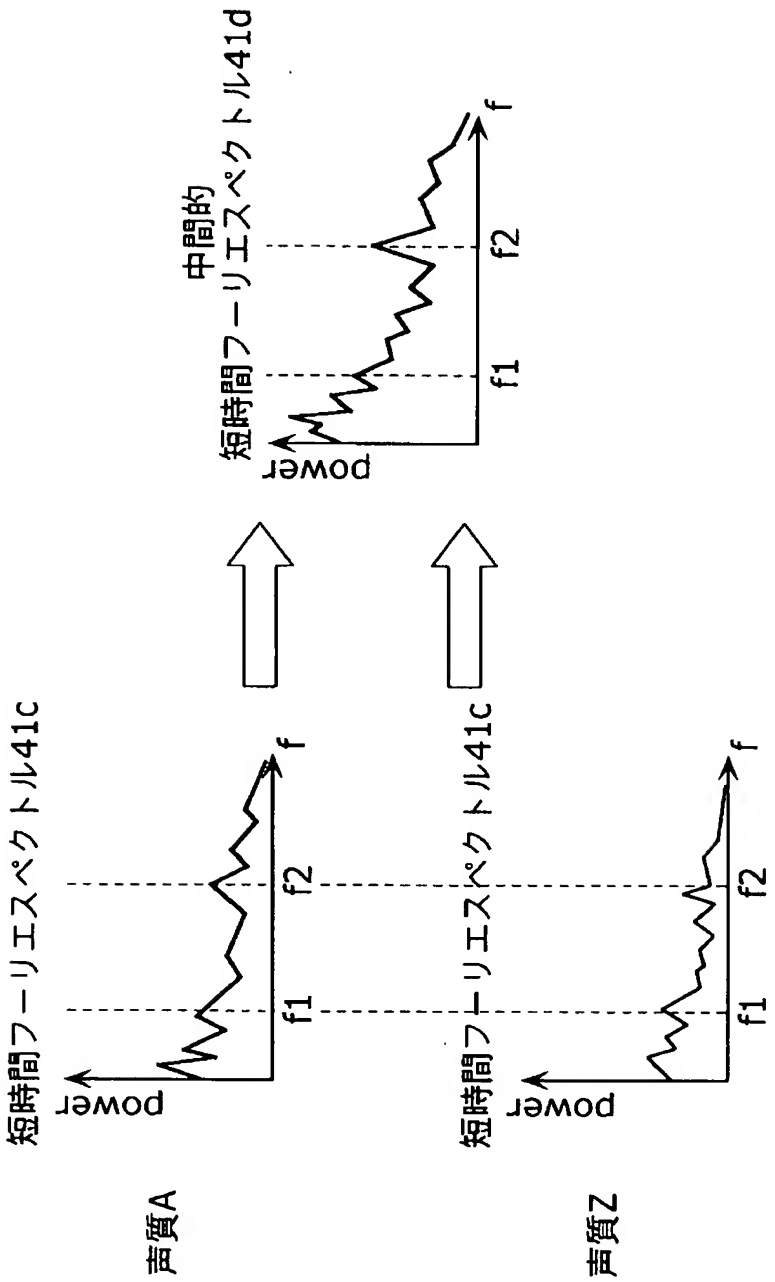
[図10]


合成音スペクトル41
短時間フーリエスペクトル41a
声質A
f
t
50
51
power
50a
50b
f
合成音スペクトル41
短時間フーリエスペクトル41a
声質Z
f
t
50
51
power
50a
50b
f

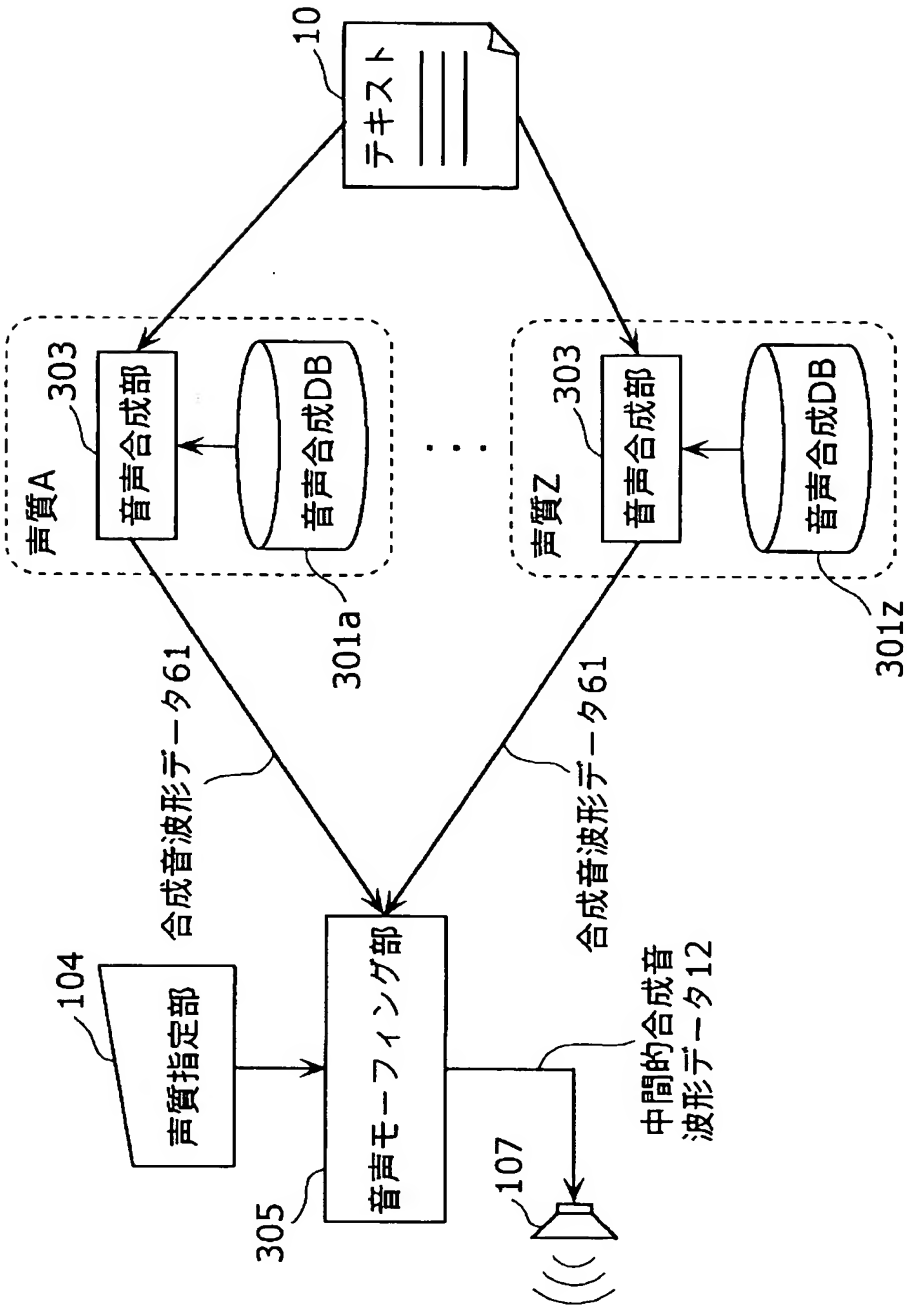
[図11]


短時間フーリエスペクトル41a
短時間フーリエスペクトル41b
声質A
power
f
50a
50b
f1
f2
短時間フーリエスペクトル41a
短時間フーリエスペクトル41b
声質Z
power
f
50a
50b
f1
f2

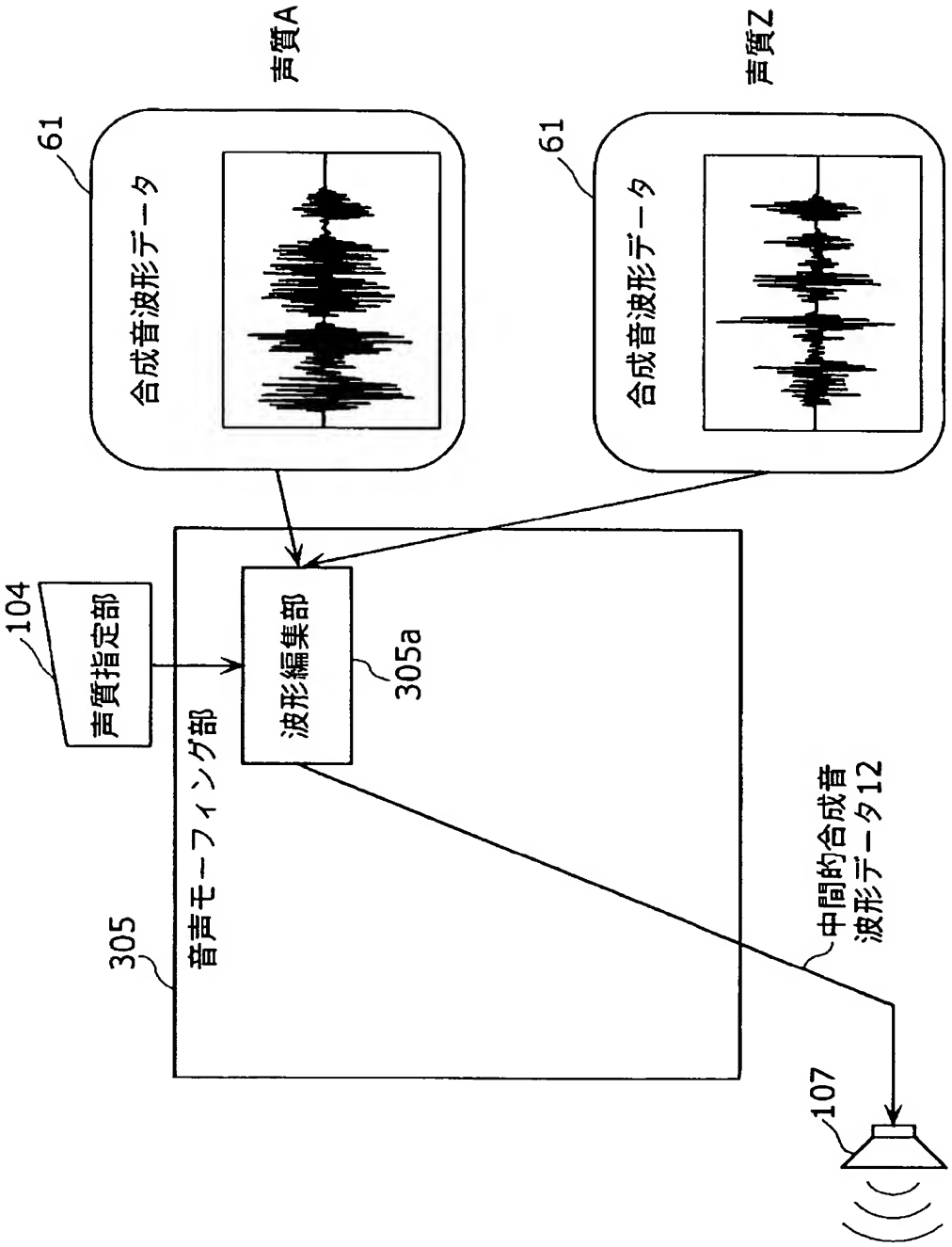
[図12]


短時間フーリエスペクトル41c
power
f
f1
f2
声質A
短時間フーリエスペクトル41c
power
f
f1
f2
声質Z
中間的
短時間フーリエスペクトル41d
power
f
f1
f2

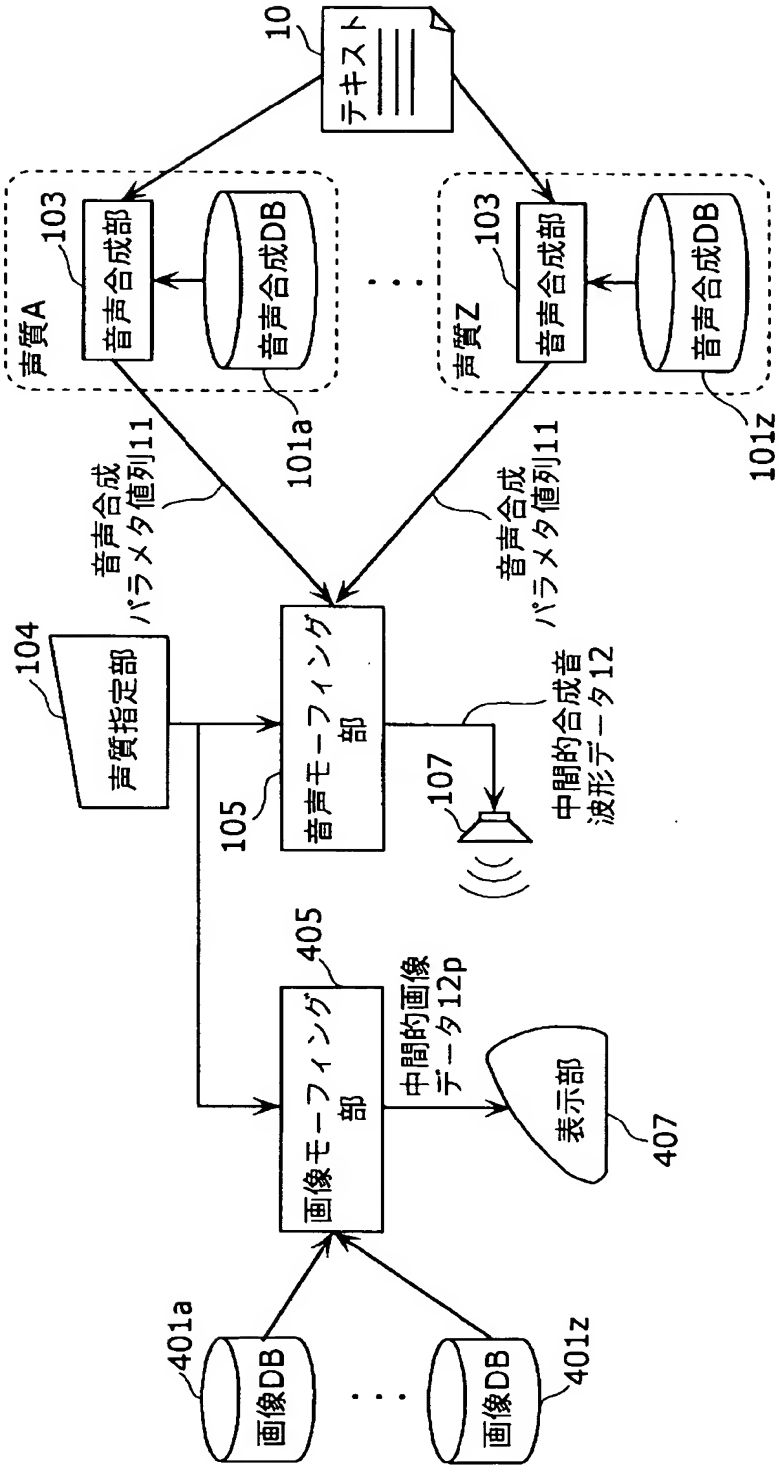
[図13]


10
テキスト
声質A
303
音声合成部
音声合成DB
301a
声質Z
303
音声合成部
音声合成DB
301z
合成音波形データ61
合成音波形データ61
104
声質指定部
305
音声モーフィング部
107
中間的合成音
波形データ12

[図14]


声質A
声質Z
61
61
合成音波形データ
合成音波形データ
104
声質指定部
305
音声モーフィング部
波形編集部
305a
中間的合成音
波形データ12
107

[図15]


10
テキスト
声質A
103
音声合成部
音声合成DB
101a
声質Z
103
音声合成部
音声合成DB
101z
音声合成
パラメタ値列11
音声合成
パラメタ値列11
104
声質指定部
105
音声モーフィング
部
107
中間的合成音
波形データ12
405
画像モーフィング
部
中間的画像
データ12p
表示部
407
401a
画像DB
401z
画像DB

[図16]